Conversation with drttol at 2005-01-21 16:45:07 on David Gucwa (aim)
(16:45:07) dr ttol: hi
(16:46:53) David Gucwa: hi
(16:50:18) dr ttol: status?
(16:50:50) David Gucwa: coming along
(16:51:04) David Gucwa: I don't expect to be done with this for several days
(16:51:46) dr ttol: okay, there is a hitch
(16:52:16) David Gucwa: okay
(16:52:40) dr ttol: because the server will be dinging the websites
(16:52:47) dr ttol: they will be all coming from the same IP
(16:52:53) dr ttol: so those websites can turn off access to that IP
(16:53:01) David Gucwa: yeah
(16:53:02) David Gucwa: they could
(16:53:03) dr ttol: we're thinking about making it client based
(16:53:09) dr ttol: can you make it a java applet
(16:53:12) dr ttol: so the users do all the importing
(16:53:19) dr ttol: and passes the arguments to a server side php script
(16:53:22) dr ttol: (so we dont give sql access out)
(16:53:36) dr ttol: so the connectu aspect is server side
(16:53:42) dr ttol: the importing aspect is client side
(16:54:27) David Gucwa: I'm trying to think if there's any easier way to get around the static IP
(16:54:33) dr ttol: okay
(16:56:10) dr ttol: proxies
(16:56:30) David Gucwa: how do those work
(16:56:59) dr ttol: communicate through proxies, but thats a hostile thing
(16:57:06) dr ttol: how hard would it be to build a java applet
(16:57:41) David Gucwa: not very, I just can't think of a good way to get the applet to send the information back to a php script
(16:58:44) David Gucwa: another concern there is that if we have it client side, rival sites could go and look at how the code works
(16:58:55) dr ttol: hm
(16:59:03) David Gucwa: and either steal it or change their own site to break it
(16:59:04) dr ttol: thats not good
(16:59:20) dr ttol: any way to do it
(16:59:27) dr ttol: protect source
(16:59:45) David Gucwa: hm
(17:00:15) David Gucwa: if the client fetched all the facebook HTML and then passed it to a php script which breaks it apart
(17:00:22) David Gucwa: I'm not sure how technically feasible that is
(17:01:14) dr ttol: we dont want thefacebook and other sites to be able to easily shut off the importer
(17:01:22) dr ttol: we'd rather play cat and mouse
(17:01:25) dr ttol: than have them just firewall us off
(17:02:37) David Gucwa: I wonder if there are third party services we can employ to fetch data for us?
(17:02:44) David Gucwa: or is that what a proxy is
(17:03:01) dr ttol: yes, proxy
(17:03:13) dr ttol: we push http data through the proxy
(17:03:25) dr ttol: but we'd need a huge list of proxies
(17:03:30) dr ttol: so why not just make a client side java app
(17:03:38) dr ttol: that just logs in and gets the html source
(17:04:14) David Gucwa: that seems like the best way to do things
(17:04:30) dr ttol: so the bulk of the code can still work
(17:04:43) David Gucwa: yeah
(17:06:18) dr ttol: ok lets do it
(17:07:29) David Gucwa: alright i'll try to set this all up for client side

Conversation with drttol at 2005-01-24 11:45:54 on David Gucwa (aim)
(11:45:54) dr ttol: hi
(11:47:57) David Gucwa: hi
(11:48:21) dr ttol: are you working?
(11:48:34) David Gucwa: yeah
(11:48:39) dr ttol: what are youw orking on?
(11:48:49) David Gucwa: setting up the proxy part of the script
(11:48:59) dr ttol: you mean the gateway thingy
(11:49:01) David Gucwa: yeah
(11:49:13) dr ttol: to allow it to discern where to send the login info to etc
(11:49:27) David Gucwa: what do you mean
(11:49:38) dr ttol: we'll have a lot of shells
(11:49:54) dr ttol: we need the script to be able to efficiently route requests to those shells
(11:50:02) dr ttol: so that those shells execute the requests
(11:50:08) dr ttol: so it appears from different ips
(11:50:22) dr ttol: you're working on the router script essentially, right?
(11:50:51) David Gucwa: I was thinking of how I'm going to do this, and the best way I could come up with is to route all requests through the same shell, then as soon as that gets blocked, it immediately moves onto a different shell.
(11:51:02) David Gucwa: if we use all shells at the same time, they could conceivably just block them all at once
(11:51:46) dr ttol: ok, but
(11:51:54) dr ttol: we're talking about 3-10 different social networks
(11:52:03) dr ttol: 2 might block
(11:52:06) dr ttol: 8 might not
(11:53:16) David Gucwa: I'm setting up a separate list for each network. They'll all start out the same at the beginning but it will know which networks have blocked which shells.
(11:53:30) David Gucwa: Actually it would probably be better to start them off in different shells just to distribute the load
(11:53:44) David Gucwa: anyway, one particular network shouldn't be seeing more than one IP from us at a time
(11:55:30) dr ttol: andyou'll be able to catch when they block us?
(11:56:09) David Gucwa: if they firewall us out, then their server just won't respond and I catch that already
(11:56:23) dr ttol: ok, or they might feed us bad info
(11:56:25) David Gucwa: if they do something funny, like changing around the text on their page when it's our IP, then I'm going to have to make changes on the fly
(11:56:26) dr ttol: etc
(11:56:35) dr ttol: yeah, i was thinking
(11:56:40) dr ttol: we need a stable case
(11:56:46) dr ttol: like an i2hub account on each network
(11:56:50) dr ttol: that we test right now, save results
(11:56:56) dr ttol: and you have it check every 5 minutes on each network
(11:57:01) dr ttol: if the results change at any point
(11:57:06) dr ttol: it means they've changed their layout
(11:57:10) David Gucwa: that's a good idea
(11:57:11) dr ttol: or feeding us bad results
(11:57:18) dr ttol: and then the script can beep you
(11:57:54) dr ttol: depending ont he severity of the error
(11:58:01) dr ttol: whether they jus firewalled us and the fix is moving to the next IP
(11:58:07) dr ttol: or whether they changed the entire layout
(11:58:17) dr ttol: we need it working asap
(11:58:24) dr ttol: so a beeper should let you know

(11:58:44) dr ttol: we'll also need to set up alerts so like if i2hub.com goes down, or connectu.com, it'll beep you
(11:59:11) dr ttol: so each node will have a test case
(11:59:11) David Gucwa: how do you set that up, I've never heard of that
(11:59:17) dr ttol: well
(11:59:29) dr ttol: set it up so it grabs the front page of i2hub.com
(11:59:41) dr ttol: if it matches the first result
(11:59:47) dr ttol: the stable case
(11:59:51) dr ttol: then its alive
(11:59:57) dr ttol: if its not reachable
(12:00:01) dr ttol: or 404
(12:00:03) dr ttol: or whatever
(12:00:14) dr ttol: if it doesnt contain a certain key word in the htm
(12:00:15) dr ttol: html
(12:00:20) dr ttol: then you know its not the real front page
(12:00:27) dr ttol: and site is either down, hacked, unreachable, etc
(12:00:42) dr ttol: then you send an email to a special beeper email with a short message
(12:00:46) dr ttol: and it'll show up on your beeper
(12:01:01) dr ttol: "i2web down"
(12:01:04) dr ttol: "cuweb down"
(12:01:08) dr ttol: etc
(12:01:28) David Gucwa: ah, it was the beeper part that I was unsure of
(12:01:34) dr ttol: ill get one as well
(12:01:51) dr ttol: but hopefully you're good enough that by the time i call you, you would hve already left a message on my voice mail ;)
(12:02:41) dr ttol: im about 1.5 hours away right now from amherst, on a 56k meeting. i was in meetings in boston all weekend
(12:02:45) dr ttol: do me a favor
(12:02:51) David Gucwa: ok
(12:02:53) dr ttol: log into thefacebook with one of those accounts
(12:03:09) dr ttol: tell me if you're able to grab email addresses from the friends listed on that account
(12:04:08) David Gucwa: Technically possible but I think it would take a long time to do, because you'd have to descend into each of the separate profile pages for each friend.
(12:04:52) dr ttol: this is what we want for that feature:
(12:05:00) David Gucwa: also that would mean a separate connection for each friend, and they'd be logging our IP each time
(12:05:05) dr ttol: a checkbox on the importer page:
(12:05:35) dr ttol: [ ] Automatically search ConnectU for my existing friends and add them to my ConnectU account
(12:05:56) dr ttol: if checked, it will descend into the profiles and try to match the emails with existing connectu members
(12:06:24) dr ttol: if there is no match, send out an invite email, and also add the email to a master database (so we dont email the same email address twice)
(12:06:49) dr ttol: what we're creating essentially is a social network spider
(12:07:42) dr ttol: how difficult will this be
(12:07:48) David Gucwa: This one person I'm looking at has 91 friends on her list. That's going to take at least 5 minutes to crawl
(12:08:00) dr ttol: really? why
(12:09:02) David Gucwa: The problem is that it doesn't show your friend's email addresses directly on your profile page, you have to go into each friend's profile, and there's a couple seconds for each one you want to load.
(12:09:33) dr ttol: call me
(12:09:35) dr ttol: 774 230 3332
(12:09:38) dr ttol: is my number right now

(12:09:44) David Gucwa: k
(12:45:15) David Gucwa: There's a hitch in the plan to crawl through the profiles in advance
(12:45:43) David Gucwa: when logged in, I can only seem to access profiles that are on my friends list or at the same school as I am
(12:47:17) David Gucwa: so I think we're going to have to spider them on the spot when people import
(12:47:24) David Gucwa: and cache them then
(12:59:32) dr ttol: wait
(12:59:42) dr ttol: atleast we can index entire schools
(12:59:43) dr ttol: with just one email
(12:59:44) dr ttol: right
(12:59:52) dr ttol: so we just need to get you 300 logins
(12:59:55) dr ttol: or so
(13:00:03) dr ttol: to access over 1 mil profiles
(13:00:48) David Gucwa: yeah
(13:01:03) David Gucwa: if you can just give me one big file with all the logins and passwords then I can use that
(13:05:07) dr ttol: can you use the one we gave you
(13:06:12) David Gucwa: I'm not sure this is going to work actually
(13:06:32) David Gucwa: I just timed the script, and it takes about 5 seconds to fetch a profile
(13:06:44) David Gucwa: for a million profiles that's 57 days to index them all
(13:06:48) dr ttol: thats fine
(13:07:33) dr ttol: we just want to get as much as we can
(13:07:47) dr ttol: i mean, if we have all of umass
(13:07:51) dr ttol: thats great
(13:07:57) dr ttol: prob only take a few hours
(13:08:33) David Gucwa: well, except I don't just have a list of ids that belong to umass
(13:08:36) David Gucwa: I have to try each one
(13:08:52) dr ttol: what do you mean
(13:09:34) dr ttol: with one umass login and password
(13:09:38) dr ttol: you're able to see the entire school
(13:09:57) David Gucwa: If I type in the id of a profile, and that person is from umass, then I can see it
(13:10:05) David Gucwa: but there's no way to know which ids are from umass
(13:10:14) dr ttol: if you click browse
(13:10:21) dr ttol: you get a list of users
(13:10:30) dr ttol: that are from you school
(13:11:17) David Gucwa: where does it say browse
(13:11:27) dr ttol: give me a login and password
(13:11:51) David Gucwa: lbowman@wellesley.edu divya
(13:12:52) dr ttol: social net
(13:12:56) dr ttol: is browse
(13:12:58) dr ttol: it is randomly generated
(13:13:01) dr ttol: so you have to check for duplicates
(13:14:39) dr ttol: see it?
(13:15:05) David Gucwa: yeah
(13:15:09) David Gucwa: let me see how fast I can grab ids from that
(13:15:13) dr ttol: k
(13:15:18) dr ttol: and that "next" link
(13:16:06) David Gucwa: next is just the same as reloading
(13:16:38) dr ttol: k
(13:20:10) David Gucwa: I can get 3 or 4 ids a second
(13:21:07) David Gucwa: How many people from umass would you say are on thefacebook? A few thousand?

(13:21:56) David Gucwa: It'll probably take ~6 hours to index all of umass's profiles
(13:22:03) David Gucwa: and cache them
(13:24:11) David Gucwa: actually that's not true, since I can only get a random list of them
(13:24:42) David Gucwa: I can probably get a large chunk of the ids pretty quickly, but towards the end it's going to be harder to get ids I haven't seen yet
(13:25:04) David Gucwa: when does this project have to be finished?
(13:28:15) dr ttol: which project
(13:28:29) David Gucwa: the whole thing, when is it going to go live
(13:28:46) dr ttol: later this week, so we want the spider to continue to do as much as possibl
(13:29:03) David Gucwa: Yeah the spider is going to be doing the bulk of the indexing
(13:29:18) dr ttol: would two spiders indexing at the same time be twice as fast
(13:29:30) dr ttol: would three? would X?
(13:30:11) David Gucwa: if we put them all on different servers. I'm not sure how much of a bottleneck bandwidth or cpu power would be if we had X spiders running on one computer
(13:30:46) dr ttol: and one central index
(13:31:47) David Gucwa: how many proxies do we have access to at the moment
(13:31:54) dr ttol: a few
(13:31:59) dr ttol: ill order more every day
(13:32:32) David Gucwa: well I'll index what I can
(13:32:43) dr ttol: ok
(13:32:55) David Gucwa: do you have a umass login that I can use
(13:33:52) dr ttol: one sec
(13:37:58) dr ttol: rdegutis@student.umass.edu
(13:38:02) dr ttol: ruth1783
(13:41:28) dr ttol: we can have all of umass indexed in 6 hours?
(13:47:31) dr ttol: ?
(13:49:46) David Gucwa: I can have a lot of it indexed in 6 hours
(13:49:57) David Gucwa: the problem is the randomness
(13:50:13) David Gucwa: like once we have 99.9% of the id's, then we're going to almost always get repeats
(13:50:22) David Gucwa: the more we have the harder it is to get new ones
(13:50:32) dr ttol: right
(13:50:39) David Gucwa: it's diminishing returns
(13:50:44) dr ttol: so get as much as you can, then move into another school
(13:50:47) David Gucwa: k
(13:50:51) dr ttol: how many ids in 6 hours
(13:51:06) David Gucwa: At least a thousand
(13:51:29) dr ttol: tomorrow, can we set it up so that we can crawl two separate schools at the same time
(13:52:07) dr ttol: does it take 5 seconds from your house?
(13:52:25) dr ttol: would it be faster if it was on i2hub's web machine
(13:52:56) David Gucwa: probably
(13:53:12) dr ttol: can you test the difference
(13:56:02) David Gucwa: yeah
(14:03:36) David Gucwa: I'm not sure this is going to work actually
(14:03:39) dr ttol: ?
(14:03:59) David Gucwa: I'm running into problems getting the social network page to show from commandline
(14:04:59) dr ttol: well, they are reachable by browser
(14:05:06) dr ttol: so its not impossible